Expert Systems with Applications 41 (2014) 7350–7360

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Proposal of a new stability concept to detect changes in unsupervised data streams

Rosane M.M. Vallim *, Rodrigo F. de Mello

ICMC, Universidade de São Paulo, Av. Trabalhador São Carlense 400, São Carlos, SP 13566-590, Brazil

ARTICLE INFO

Article history:
Available online 21 June 2014

Keywords:
Data streams
Unsupervised change detection
Surrogate stability
Surrogate data

ABSTRACT

Learning from continuous streams of data has been receiving an increasingly attention in the last years. Among the many challenges related to mining data streams, change detection is one topic frequently addressed. Being able to determine whether or not data characteristics are changing along time is a major concern for data stream algorithms, be it on the supervised or unsupervised scenario. The unsupervised scenario is particularly relevant due to many practical applications do not provide target labeling information. In this scenario, most of the strategies induce consecutive models over time and compare them in order to detect data changes. In this situation, model changes are assumed to be a consequence of data modifications. However, there is no guarantee this assumption is true, since those algorithms do not rely on any theoretical background to ensure that model divergences truly indicate data changes. The need for such theoretical framework has motivated this paper to propose a new stability concept to establish bounds on the learning abilities of unsupervised algorithms designed to detect changes on data streams. This stability concept, based on the surrogate data strategy from time series analysis, provides learning guarantees for online unsupervised algorithms even in case of time dependency among observations. Furthermore, we propose a new change detection algorithm that meets the requirements of this stability concept. Experimental results on different synthetical scenarios illustrate how the stability concept proposed in this paper is applied to detect changes in unsupervised data streams.

1. Introduction

Data Stream Mining is an active area of research that is concerned with the development of algorithms capable of learning models from data streams. Data streams are ordered, infinite sequences of data, that become available along time (Gama & Rodrigues, 2007). Due to its infinite nature, researchers usually assume the probability distribution responsible for generating a stream is not fixed nor stationary, consequently data characteristics evolve over time. This evolving aspect has motivated several studies to design algorithms to detect when data is actually changing, allowing for efficient and effective model reinduction in the presence of new data behavior.

In the supervised learning scenario, the most common strategies for detecting data changes monitor some performance measures of the induced model, such as accuracy or precision (Gama, Medas, Castillo, & Rodrigues, 2004). If these measures fall below a stablished threshold, then the current model is considered outdated and, therefore, no longer useful for making predictions about data. A change in the data distribution is then issued, and the model is reinduced using new data. This strategy gives a fair indication on changes, however it requires labeled examples. Unfortunately, most of real-world applications only provide unsupervised data, meaning there is no *a priori*knowledge to consider when inducing models.

In this context, many researchers have been designing clustering techniques to approach the unsupervised scenario, as well as measures to monitor clustering evolution along time in an attempt to detect data changing behavior (Albertini & Mello, 2010; Marsland, Shapiro, & Nehmzow, 2002; Vallim, Filho, de Mello, & de Carvalho, 2013). These strategies assume that changes observed in the induced models indicate changes in data characteristics. However, those strategies have no formal guarantee due to algorithm parameters can lead to model adaptations that may not correspond to data modifications.

The lack of learning guarantees for unsupervised scenarios has motivated Carlsson and Memoli (2010) to propose a stability concept for unsupervised batch learning, which is formalized in terms of model divergences when input data is subject to order perturbations. According to this concept, an algorithm is proven to be stable

* Corresponding author. Tel.: +55 16 9607 9831.
E-mail addresses: rosane.maffei@gmail.com (R.M.M. Vallim), mello@icmc.usp.br (R.F. de Mello).

http://dx.doi.org/10.1016/j.eswa.2014.06.031
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

if it produces the same model regardless of permutations in data ordering. The theoretical foundation of such work is developed assuming batch learning and considering data is independent and identically distributed (i.i.d.).

When considering a data stream, however, an infinite number of observations may be generated, which require incremental model induction as opposed to the traditional multiple pass used in batch learning. Moreover, while data independency could be assumed for some streams, in many others time information cannot be disregarded, i.e., data dependencies are important and removing this information would completely change the problem characteristics or the phenomenon represented by data (Ceccherini, Gobron, & Migliavacca, 2014; Ferlay, Bray, Steliarova-Foucher, & Forman, 2014; Ferlay et al., 2008; Takahashi, Akiniwa, & Narita, 2001). Therefore, since not all data is previously available for learning, and it may contain time dependencies, an stability concept purely based on data permutations will fail when applied on the data stream domain.

In an attempt to provide learning guarantees for the unsupervised data stream scenario, this paper proposes a new stability concept that assumes observations must be processed as long as they are collected and are not necessarily i.i.d. In order to design this new concept, we first assume observations in a stream present an inherent order that cannot be disregarded. Based on this, we further assume that a data stream can be seen as a time series, which we believe is fair enough for the stream domain because observations may have different levels of dependency over time, including none. Since independency among observations cannot be straightly assumed when considering a time series, ordering permutations, such as the one proposed by Carlsson and Memoli (2010), cannot be used as a source of data perturbation when analyzing the stability of data-stream algorithms.

With this in mind, the stability concept here proposed is based on model divergences when surrogate data (Theiler, Eubank, Longtin, Galdrikian, & Doynefarmer, 1992) is used as input for the learning algorithm. Surrogate series are new data observations generated from taking the original series as input. These new observations maintain certain characteristics of the original series, such as frequency and amplitude. Our stability concept, called hereafter surrogate stability, states that an algorithm is stable if it produces the same models for both the original and the surrogate series. In analogy with the data-stream scenario, if a stream maintains the same properties along time, models induced at different time instants must be equal. In this situation, data collected in two consecutive time instants correspond to the original time series and its surrogate, respectively. Since the surrogate stability do not alter the order of the observations, possible time dependencies among observations are taken into account, making this stability concept suitable for the data stream domain.

Furthermore, this work shows that algorithms producing a Power Spectrum (PS) graph as the data stream model are stable according to the surrogate stability concept. A new change detection algorithm is then proposed, which compares PS graphs in two consecutive time-windows of data, and issues changes when these models diverge from each other. The idea behind the algorithm is that if no change happens in two consecutive windows, the data inside each window presents very similar properties. In this scenario, the sub-series contained in the most recent window can be seen as a surrogate of the sub-series contained in the previous window. On the other hand, if a change happens in data characteristics, we assume frequencies and amplitudes will also be modified, and the most recent sub-series can no longer be seen as a surrogate of the previous one. Results confirm that the proposed algorithm correctly indicates when the data stream changes, guaranteeing the detection is a consequence of data modifications and not due to influences produced by model parametrization such as in previous approaches (Albertini & Mello, 2010; Marsland et al., 2002; Vallim et al., 2013; Bifet & Gavaldà, 2007).

This paper is organized as follows. In Section 2, we present related work in the area of change detection as well as other stability concepts to provide learning guarantees. Next, Section 3 provides the formal background used to develop the surrogate stability concept, the definition of the concept itself, as well as the new stable change detection algorithm for unsupervised data streams. Experimental studies to verify the surrogate stability concept and the change detection algorithm proposed are detailed in Section 4. Such experiments explore synthetic scenarios, allowing to confirm whether changes were detected when they should. Section 5 brings a further analysis on the results observed in these experiments. Finally, concluding remarks and future work are presented in Section 6.

2. Related work

2.1. Unsupervised change detection in data streams

Several different strategies have been proposed in the datastream literature to deal with the change detection problem, ranging from statistical algorithms based solely on summary measures such as mean and variance, to more advanced techniques using clustering and novelty detection.

Page (1954) approaches change detection in an incremental fashion by monitoring the cumulated difference between the observations in the stream and their mean. This method, called Page–Hinkley Test (PHT), issues a change if the monitored statistic falls below a user-defined threshold value. Adaptive Windowing (ADWIN) (Bifet & Gavaldà, 2007) also uses mean to detect changes in data. However, ADWIN applies time-windows and successively divides this window in two subwindows, comparing their means and issuing a change if the difference is greater than a threshold. Both approaches fail in presence of amplitude and frequency modifications over the data collection. For example, consider two different scenarios: (i) a stationary stream in which the data mean does not modify over time, however the data frequency does; (ii) a weakly-stationary stream, in which data variance may change over time, however the mean is always the same. In both situations, change detection would only be possible if: (i) the algorithm gives a greater relevance to recent observations in case of PHT, making it forget most of past information; or (ii) if the algorithm reduces the window length to detect changes in case of ADWIN. However, those attempts would make both algorithms map any small perturbation as a change what is not desirable.

Recently, change detection has also been approached applying incremental clustering algorithms and evaluating model divergences at different time instants, i.e., considering new observations to adapt models. The Grow When Required (GWR) neural network (Marsland et al., 2002) issues changes when new neurons are added to the model or when a neuron that has not been used recently is activated. This algorithm, therefore, can issue changes based on a single new observation, for example an outlier. Another neural-network-based algorithm is the Self-Organizing Novelty Detection Neural Network (SONDE) (Albertini & Mello, 2010), which uses Shannon's entropy to quantify the level of novelty introduced in the model after receiving a new observation. Therefore, the algorithm considers that model divergences correspond to data changes. Recently, Vallim et al. (2013) proposed M-DBScan, a change detection algorithm that uses a density-based clustering method together with an entropy measure to estimate model divergences. M-DBscan design considers that a change should be seen as a sequence of novel events, therefore this approach is more robust to outliers than previous methods. Despite of the good